

The relationship between stream placement and teachers' judgements of pupils: Evidence from the Millennium Cohort Study

Tammy Campbell*
London School of Economics

Abstract

This paper tests the hypothesis that stream placement influences teacher judgements of pupils, thus investigating a route through which streaming by 'ability' may contribute to inequalities. Regression modelling of data for 800+ 7-year-olds taking part in the Millennium Cohort Study examines whether teachers' reported perceptions of 'ability and attainment' correspond to the stream in which a pupil is situated. Children with similar characteristics, who perform equivalently on recent, independent, salient cognitive tests, and who have equal prior attainment, are compared. As predicted, stream level is associated with teachers' perceptions. The hypothesis that there is a relationship from stream placement to teacher judgement is supported.

Keywords: primary education; streaming; perceptions; judgements; assessments; Millennium Cohort Study

Introduction

Research and policy context

Streaming, the practice of grouping by 'ability' all pupils within a school year cohort, has increased in English primary schools since the turn of the century. In the space of a decade, estimates of the prevalence of the practice have grown from less than 2 per cent of all primary children in 1999 (Hallam *et al.*, 2003) to nearly 18 per cent of Year 2 pupils in 2008 (Campbell, 2013).

The resurgence of streaming has been backed by a government-sanctioned push towards various forms of 'ability'-grouping (Boaler, 1997; Conservative Party, 2007; DCSF, 2008; DfE, 1992; DfE, 2010; DfES, 2005). This is despite the majority of the available evidence indicating that early grouping neither raises overall average attainment nor leads to greater parity in opportunity or achievement (Dunne *et al.*, 2007; Slavin, 1990; Higgins *et al.*, 2014). International research by the OECD, for example, has suggested that '[e]arly student selection has a negative impact on students assigned to lower [streams] and exacerbates inequities, without raising average performance' (OECD, 2012: 10). Kutnick *et al.* (2005: 12) reviewed a mostly British literature and concluded that pupil ability groupings 'appear to have replicated the achievement spectrum that they were designed to reduce'. The Education Endowment Foundation – sponsored by the Department for Education to gather and evaluate evidence to inform policy – conclude their recent synthesis by recommending that streaming 'does not appear to be an effective strategy' (Education Endowment Foundation, 2017: 1).

* Email: t.campbell1@lse.ac.uk

©Copyright 2017 Campbell. This is an Open Access article distributed under the terms of the Creative Commons Attribution Licence, which permits unrestricted use, distribution and reproduction in any medium, provided the original author and source are credited.

Francis et al. (2017: 1) have argued that the apparent lack of research impact on policymaking around streaming can be explained by ‘cultural investments in discourses of “natural order” and hierarchy’. They propose that there is a historically underpinned, ‘common sense’ notion that children are of different ‘types’, and can, fairly and without detrimental impact, be sorted into streams. Corresponding assumptions of innate and immutable difference are apparent throughout the policy documents that have accompanied the increase in streaming: the 2005 Education White Paper (DfES, 2005: 20), for example, delineated pupils as ‘gifted and talented’, ‘struggling’ or ‘just average’ – while the 2016 White Paper (DfE, 2016: 98) continues to talk of ‘[s]treetching both the lowest-attaining and *most able*’ (my italics).

As noted, the sufficiency of these notions and assumptions as a rationale for fair and effective streaming has repeatedly been challenged by the academic literature. As well as indicating that the practice is inefficient and inequitable in its impact on children’s eventual attainment, studies have demonstrated disparities in ‘ability’ grouping placements themselves that reflect wider societal inequalities (Ansalone, 2003; Boaler, 1997; Boaler et al., 2000; Kutnick et al., 2005, Wiliam and Bartholomew, 2004). The most recent UK evidence (Hallam and Parsons, 2013) suggests, for example, that, even after controlling for prior attainment and cognitive aptitude, low-income primary school pupils are disproportionately often placed in the lowest streams, along with children with less-educated parents. There are, moreover, some indications of disproportionality by ethnicity. These disparities in placement according to pupil characteristics seem, therefore, to indicate that, rather than providing a structure that facilitates a pupil’s progress through some ‘natural’ trajectory, streaming may intervene actively to impose a structure that can entrench between-group differences.

Routes through which streaming may contribute to inequalities

Through what processes might streaming contribute to differentiation in children’s academic development? Research has proposed several mechanisms. First, there is evidence that a pupil’s own self-concept, perceptions and behaviours can be influenced by the group to which they are assigned (Ansalone, 2003; Boaler, 1997; Croizet and Claire, 1998; Kutnick et al., 2005; Reay, 2006; Shih et al., 2005; Steele and Aronson, 1995; Yopyk and Prentice, 2005). Second, studies suggest that educational opportunities and quality of teaching may differ according to stream placement, with the progress of children in upper groups being facilitated to a higher level than those placed at the bottom of the hierarchy (Ansalone, 2003; Boaler, 1997; Kutnick et al., 2005). Third, evidence indicates that stream placement might influence the perceptions and expectations that class teachers hold of their pupils. The wider literature demonstrates that teachers can (consciously or unconsciously) label and stereotype children based on a variety of characteristics (Burgess and Greaves, 2009; Campbell, 2015; Hansen and Jones, 2011; Hansen, 2016; Reeves et al., 2001; Thomas et al., 1998). In particular, there is evidence that teachers formulate and act upon expectations of pupils according to the level of their academic group placement (Ansalone, 2003; Boaler, 1997; Boaler et al., 2000; Ireson and Hallam, 1999; Rubie-Davies, 2010). Assigned stream level may therefore affect teacher perceptions of their whole class and of each pupil within the class.

This is crucial, not least because there are well-established relationships between teacher judgements and pupil attainment. From the experimental explorations of Rosenthal and colleagues in the 1960s (Rosenthal and Jacobson, 1968) to the present, a solid body of evidence has built that suggests that teacher beliefs about, and expectations of, their pupils can influence pupils’ achievement: ‘when teachers believe ... their students [are] very able [they interact] with them in ways which promote ... their academic development’ (Rubie-Davies, 2010: 121; see also

Alvidrez and Weinstein, 1999; Brophy and Good, 1970; Good, 1987; Miller and Satchwell, 2006; Rubie-Davies et al., 2014).

As most academic attainment at the primary level is currently judged and assessed by teachers, processes of perceptual bias, and their potential influences on pupils, are more important than ever. If stream placement affects a teacher's sense of a child, this could impact meaningfully upon the child's education.

The current study: Rationale and hypothesis

Teacher judgements are therefore the focus of this paper. Analyses seek to explore whether they can directly be implicated in the process through which streaming may contribute to differentiated outcomes. While some previous studies have investigated the relationships between stream placement and teachers' views of pupils, most have been small-scale, and explicit controls for the impact and mediation of the many factors and processes that may confound any direct associations have been sparse (Blatchford et al., 2010; Ireson and Hallam, 1999; Kutnick et al., 2006). There is a dearth of up-to-date UK research, particularly in the primary sector – presumably due, in part, to the fact that the documented resurgence of streaming among young pupils has arisen fairly rapidly, since the turn of the century (Hallam and Parsons, 2013).

Only lately have studies begun to exploit the potential of emerging quantitative data, primarily from the Millennium Cohort Study (MCS), to identify the possible effects of different ability-grouping practices on pupil progress and attainment. Parsons and Hallam (2014), for example, compare the trajectories of children who are streamed to their counterparts in schools employing 'mixed-ability' teaching.

The current paper therefore continues to build upon analyses of the MCS, using a sample of pupils in early primary school in England. It accounts for a broad variety of factors that may confound connections between stream placement and teacher perceptions, exploring the data using detailed regression modelling. Analyses here hope more definitely to isolate associations between stream placement and perceptions, and to test the hypothesis that teacher judgements of pupils are influenced by the stream to which a pupil is allocated.

Methodology

Sample and data

The MCS is a longitudinal sample including 11,695 babies born in England around the turn of the century. The children and/or their families have been interviewed six times to date: within the child's first year (2001), then at ages 3 (2004), 5 (2006), 7 (2008), 11 (2012) and 14 (2015) (Hansen, 2012).

In 2008, an English subsample of MCS children's teachers responded to a survey asking about their perceptions of the child's attainment, of their behaviours, and for details of the grouping structures within their schools. The teachers of 5,598 children participated, meaning that data are available for 63 per cent of the 8,887 children comprising the main wave-four sample (Johnson et al., 2011). Of the sample pupils in state schools, 914 (17.5 per cent) are reported as being streamed, and data on stream placement itself is available for 882 English, 7-year-old, singleton pupils within this group, of whom 851 also have information on teacher judgements (see University of London, 2008; 2011a; 2011b; 2012a; 2012b for data source references). The MCS sample pupils for whom stream placement information is available differ only minimally from those English, singleton, state school MCS children who are reported as not being streamed, according to a number of key characteristics (see Campbell, 2014: 40).

Unless otherwise stated, all estimates are weighted for the MCS's design features and for attrition to the main wave-four sample, as per Mostafa (2013), and are produced using Stata version 14. Listwise deletion is used to maximize sample sizes, so there is some variation in numbers.

Outcome variables

The main regression analyses undertaken in this paper use as outcomes teacher survey reports of each pupil's 'ability and attainment'. These judgements are chosen on the basis that they were provided independently, as part of the MCS, so are not prone to the distortion and biases that skew assessments made within the high-stakes schools accountability system, such as key stage teacher assessments (Bradbury, 2011; Campbell, 2015; Harlen, 2007). A sensitivity check using children's Key Stage 1 scores is, however, performed additionally; findings are congruent and are detailed in the results section.

During the MCS teacher survey, respondents were asked to 'rate ... the study child's ability and attainment ... in relation to all children of this age'. Teachers could choose to define a pupil as: 'well above average', 'above average', 'average', 'below average' or 'well below average'. Ratings were recorded for teacher perceptions of the child's 'ability and attainment' across seven domains: speaking and listening, reading, writing, science, maths and numeracy, physical education, information and communication technology, and expressive and creative arts. The first analysis in this paper allocates each sub-response a score of 1 to 5 (where 1 represents 'well below average' and 5 'well above average'), and sums these scores to represent one 'overall' rating, ranging from 7 to 35 (mean = 22; SD = 5.3). This seeks to represent each teacher's general judgement of a pupil's capabilities, and analysis using this outcome is modelled using linear regression.

Among the 851 sample pupils, responses for each domain are, in the main, highly correlated with this overall summed total (see Table I). Judgements of ability in physical education and in arts are less strongly related to the total and to judgements in each other subject, suggesting some delineation between teacher perceptions of performance in 'academic' and 'non-academic' domains. Therefore, the summed total including all subjects is used for the main analysis, and sensitivity checks excluding judgements on physical education and arts are also carried out (scale 5–25).

Table I: Correlations between summed teacher judgement and judgements in each individual domain

	Overall ability	Reading ability	Writing ability	Science ability	Maths ability	PE ability	ICT ability	Arts ability
Overall ability	1.00							
Reading ability	0.90	1.00						
Writing ability	0.91	0.87	1.00					
Science ability	0.90	0.78	0.78	1.00				
Maths ability	0.89	0.80	0.80	0.80	1.00			
PE ability	0.66	0.42	0.47	0.51	0.48	1.00		
ICT ability	0.84	0.68	0.68	0.73	0.70	0.60	1.00	
Arts ability	0.74	0.56	0.59	0.60	0.52	0.57	0.62	1.00

N = 851 (unweighted). All estimates weighted for survey design and attrition to main wave-four survey.

Further analyses are performed separately for judgements of reading and of maths ability respectively (here, the scale is 1–5), using ordered probit modelling. Three main survey-reported teacher judgements of ‘ability and attainment’ are therefore used as outcomes:

- (1) Aggregated overall judgement (range: 7–35) – modelled using linear regression.
- (2) Judgement of reading ability (range: 1–5) – modelled using ordered probit regression.
- (3) Judgement of maths ability (range: 1–5) – modelled using ordered probit regression.

Key predictor variable: Stream placement

The key predictor in modelling against all outcomes is a pupil’s stream placement (delineated in the survey as ‘top’, ‘middle’ or ‘bottom’), as reported by their teacher. Streaming is defined in the questionnaire as ‘group[ing] children in the same year by general ability and they are taught in these groups for most or all lessons’. In the sample of 851 pupils, 41 per cent are reported as being in the top stream, 31 per cent in the middle stream and 28 per cent in the bottom stream.

Key controls: Recent cognitive test scores

Very shortly before children’s teachers were contacted for their survey, the MCS pupils were visited in their homes by interviewers who administered three separate cognitive tests. The mean time lag between pupil cognitive tests and teacher survey was 3.8 months. Scores on the tests provide key controls to teacher judgements, allowing analyses of whether children who perform equivalently, but who are placed in different streams, are judged differently by their teachers.

The first of the tests is the British Ability Scales Word Reading Test, designed to assess children’s English reading ability. The ability score (a scaled but not otherwise standardized score) is utilized (see Hansen, 2012). Secondly, performance on the Progress in Mathematics Test is included. This test is designed to measure pupils’ mathematical ability across use of numbers, shapes and skill in data handling. The shortened version used in the MCS entailed routing to sections of varying difficulty levels (Rasch scaling converted the raw scores to a count score equivalent to that which would be attained were the full test completed) (see Hansen, 2012) – this scaled score is used. Lastly, the British Ability Scales Pattern Construction Test (PCT) is incorporated. This test has been developed to provide an indication of overall cognitive aptitude and, as with the Word Reading Test, the ability score is used for modelling.

Scores for all three tests are used in as ‘raw’ a form as possible (weighted/scaled only for question difficulty/routing/selection), and are not otherwise standardized or modified. This means that each simply represents a child’s manifest performance when completing that test on the given day. As children took the tests at slightly different ages within the MCS fieldwork, and because the lags between tests and teacher survey vary slightly, both pupil age at cognitive tests and pupil age at teacher survey are controlled for in all analyses, unless otherwise stated.

Figures 1, 2 and 3 illustrate the distribution of scores on the three cognitive tests for sample pupils situated in each stream.

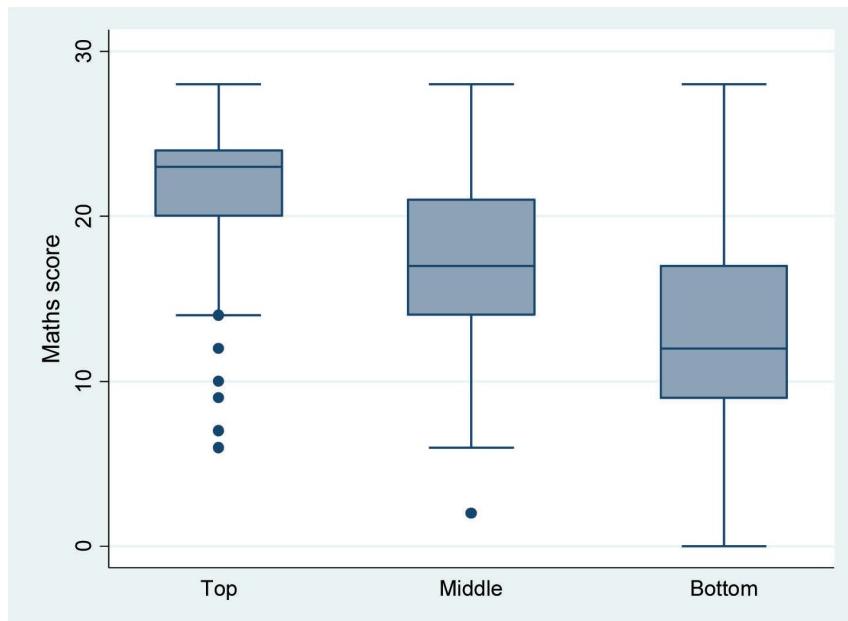


Figure 1: Distribution of Progress in Maths scores: sample pupils across streams

n = 840; Mean for all pupils = 18.2. Line represents median, box represents 25th and 75th percentiles (Q1 and Q3, respectively), whiskers represent $Q3 + 1.5(Q3 - Q1)$ / $Q1 - 1.5(Q3 - Q1)$.

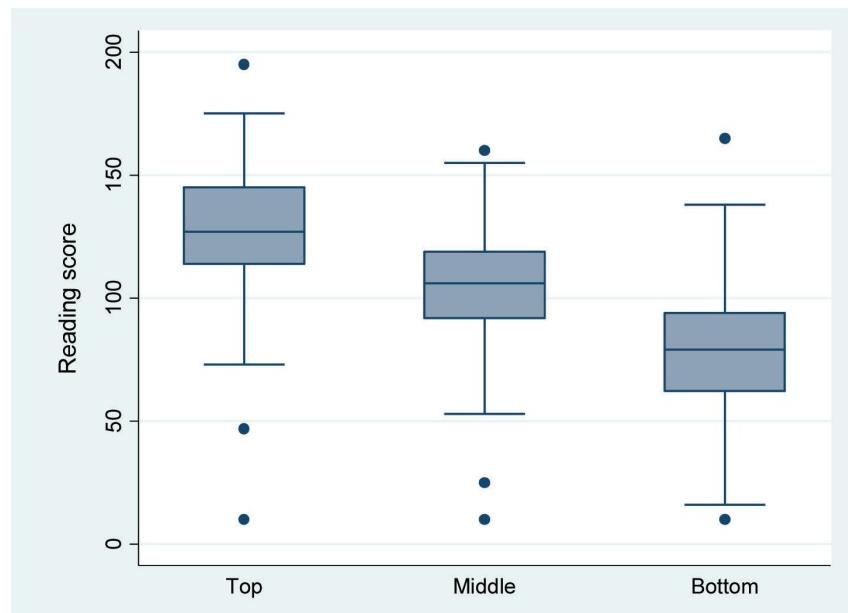


Figure 2: Distribution of Word Reading scores: sample pupils across streams

n = 837; Mean for all pupils = 108.5. Line represents median, box represents 25th and 75th percentiles (Q1 and Q3, respectively), whiskers represent $Q3 + 1.5(Q3 - Q1)$ / $Q1 - 1.5(Q3 - Q1)$.

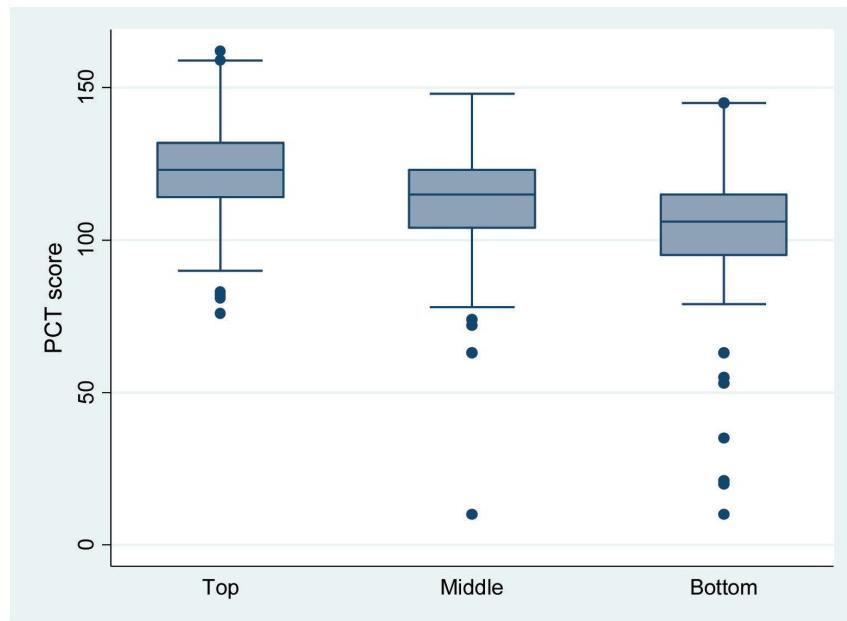


Figure 3: Distribution of PCT scores: sample pupils across streams

n = 835; Mean for all pupils = 114.6. Line represents median, box represents 25th and 75th percentiles (Q1 and Q3, respectively), whiskers represent $Q3 + 1.5(Q3 - Q1)$ / $Q1 - 1.5(Q3 - Q1)$.

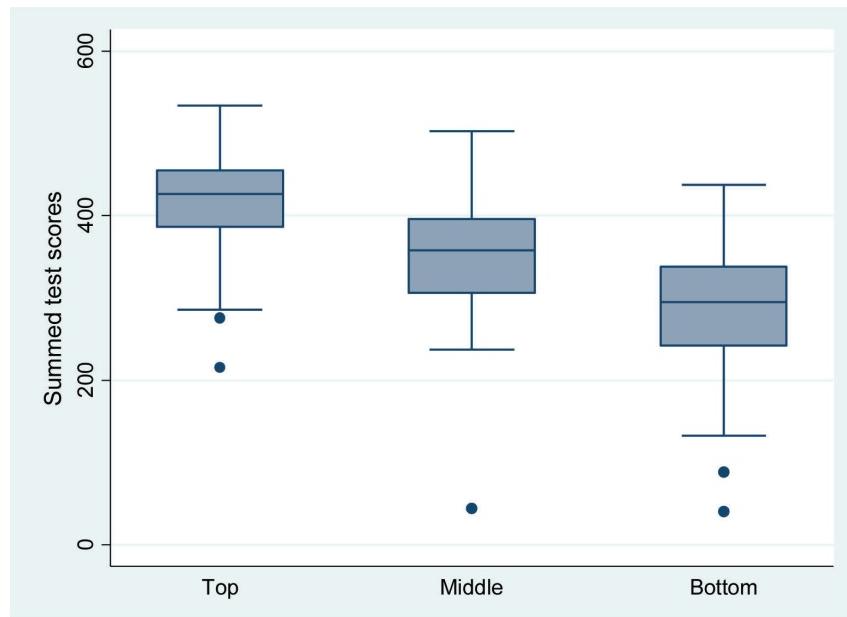


Figure 4: Distribution of summed test scores: sample pupils across streams

n = 829; Mean for all pupils = 366.6. Line represents median, box represents 25th and 75th percentiles (Q1 and Q3, respectively), whiskers represent $Q3 + 1.5(Q3 - Q1)$ / $Q1 - 1.5(Q3 - Q1)$.

While there is variation between streams, with pupils in the higher groups scoring better on average in all the tests, there is also an overlap between groups: some children who score equivalently on the cognitive tests are situated in different streams. Most overlap is apparent in PCT scores – notable given that the PCT is intended to measure ‘overall’ cognitive ability, just as overriding stream placement is intended to reflect ‘general’ ability across subjects. Figure 4 shows the distribution of each child’s combined cognitive test score across streams when the three scores are summed together and equally weighted, to provide an alternative composite representation of performance on the particular tasks undertaken. Again, there is an overlap of similarly scoring children between streams.

Table 2: Covariates

Pupil and family characteristics	Child’s behaviour/perceptions of their behaviour	Child’s special educational needs	Teacher characteristics
Pupil gender	Strengths and difficulties questionnaire (SDQ), parent-report: emotional subscale (at age 5)	Teacher-report of whether child has ‘ever been recognized as having Special Educational Needs’ (at age 7)	Gender
Pupil month of birth	SDQ parent-report: conduct subscale (at age 5)		Number of years teaching
Pupil ethnicity	SDQ parent-report: hyperactivity subscale (at age 5)		Number of years teaching at this school
Pupil’s family’s income level (at age 7)	SDQ parent-report: peer problems subscale (at age 5)		
Pupil’s main parent’s highest qualification (at age 7)	SDQ parent-report: prosocial subscale (at age 5) SDQ teacher-report: emotional subscale (at age 7) SDQ teacher-report: conduct subscale (at age 7) SDQ teacher-report: hyperactivity subscale (at age 7) SDQ teacher-report: peer problems subscale (at age 7) SDQ teacher-report: prosocial subscale (at age 7)		
	Teacher-report of overall problems with emotions, behaviour, concentration or social skills (at age 7)		

A comprehensive collection of controls, accounting for factors that have been indicated by the previous literature as potentially confounding the relationships between stream placement and

teacher perceptions, or as related to pupil performance, are also included in modelling. Table 2 outlines these variables, and descriptive information on their distribution across stream levels is available in Campbell (2014), along with information on the rationale for inclusion of each. Key to note is an unequal distribution by characteristic across streams that has similarly been described in other analyses of the MCS data (Parsons and Hallam, 2014; Campbell, 2014): boys, for example, tend more often to be placed at a lower level, along with summer-born children (who are relatively younger within cohort), pupils from low-income families and those whose parents are educated to a lower level.

Teacher perceptions of pupils may be influenced by what they know about a pupil's prior attainment, and by judgements conveyed by other staff within their school. In addition, prior attainment may have been influential in determining the stream to which a child is allocated, while also being predictive of a child's current performance. Correspondingly, Table 3 indicates an association between Foundation Stage Profile (FSP) score, assigned two years previously by the class teachers who taught the pupils' reception groups when they were 5, and stream placement at age 7. FSP score is therefore added as an additional control in modelling.

Table 3: Mean total FSP score at age 5^(a)

	Top stream	Middle stream	Bottom stream
FSP total score (range 0–117)	98.1	83.6	69.1

(a) All estimates weighted for survey design and attrition to main wave-four survey. N = 774 (unweighted).

Modelling

All analyses combine the key predictor variable (stream placement) with the controls detailed above, and regress these multiple predictors on each of the three measures of survey-reported teacher judgement. Controls are added through cumulative model specifications, in order to highlight the potential moderating effect of each set of factors, and Table 4 describes each respective specification.

Table 4: Cumulative specifications for models with each survey-reported teacher judgement as outcome

Specification	Predictors	Outcome
One	Stream placement Maths Test score Reading Test score Pattern Construction Test score Age at cognitive tests Age at teacher survey	Survey-reported teacher judgements of 'ability and attainment', summed (range 7–35; linear regression) or Survey-reported teacher judgement of maths 'ability and attainment' (range 1–5; ordered probit regression)
Two adds...	Pupil gender Pupil month of birth Pupil ethnicity Pupil's family's income level Pupil's main parent's highest qualification (age 7)	or Survey-reported teacher judgement of reading 'ability and attainment' (range 1–5; ordered probit regression)

Specification	Predictors	Outcome
Three adds...	Age 5 parent SDQ: emotional Age 5 parent SDQ: conduct Age 5 parent SDQ: hyperactivity Age 5 parent SDQ: peer Age 5 parent SDQ: pro-social Age 7 teacher SDQ: emotional Age 7 teacher SDQ: conduct Age 7 teacher SDQ: hyperactivity Age 7 teacher SDQ: peer Age 7 teacher SDQ: pro-social Teacher overall judgement of pupil behaviour	
Four adds...	Foundation Stage Profile total score (banded)	
Five adds...	Any diagnosis of special educational need	
Six adds...	Teacher gender Teacher years teaching Teacher years teaching at this school	

Chronology, support for, and assumptions behind modelling strategy

For modelling to test the hypothesis of a directional relationship from stream placement to teacher perceptions, and to rule out the possibility of reverse causality, it is necessary, firstly, that stream placement should precede teacher judgement, and secondly, that the judging teacher should not have been instrumental in determining placement.

That the first is the case rests on the timing of the MCS fieldwork and on a related assumption that cohort-wide stream placement would have been established at the beginning of Year 2, and altered little in the year that followed. Children's stream levels would then be in place before teachers reported their perception during the survey, which took place predominantly towards the end of the same academic year (Huang and Gatenby, 2010). Teachers participating in the MCS are presumed therefore to provide details of each child's established stream placement, which, crucially, has preceded their judgement of the child as provided in the same questionnaire.

The second supposition that the respondent class teacher who provides survey judgement should not have allocated the MCS pupil to their stream level, is suggested both by the nature of streaming itself and by reviews of evidence on school organizational practices. As streaming takes place at the whole-year level, placement may officially be determined by some combination of performance in previous years, assessments by previous years' teachers, pre-established placements and/or school-based test performance (Blatchford et al., 2010; Kutnick et al., 2005; Kutnick et al., 2006). In addition, as noted, drivers other than those officially stated – such as children's characteristics – seem also to be tacitly influential. Once streams have been decided upon, each set of pupils may be allocated to one of the year group's assigned class teachers – meaning that the class teacher is unlikely to be the key decision-maker determining allocations. Note that this contrasts with the probable processes behind other types of 'ability'-grouping, such as within-class grouping, where the class teacher is likely to be a key decision-maker.

Results

Table 5 presents key results for each model specification, where the outcome is summed survey-reported teacher judgement. It indicates a relationship between pupils' stream placements and their teachers' judgements of their 'ability and attainment'. Even at specification six, controlling for all potential confounders, being in the top stream is associated with overall teacher judgements of 'ability and attainment' (range: 7–35; SD: 5.3) 2.6 points higher ($p < .001$), and being in the bottom stream associated with judgements 1.7 points lower ($p < .001$).

Table 5: Difference in summed survey-reported teacher judgement of 'ability and attainment' according to pupils' stream placement^{(a)(b)}

	Spec 1	Spec 2	Spec 3	Spec 4	Spec 5	Spec 6
Top stream	3.157*** (0.286)	2.874*** (0.274)	2.661*** (0.260)	2.586*** (0.253)	2.611*** (0.250)	2.569*** (0.258)
(Middle stream)	0 (.)	0 (.)	0 (.)	0 (.)	0 (.)	0 (.)
Bottom stream	-2.702*** (0.327)	-2.384*** (0.328)	-1.964*** (0.318)	-1.897*** (0.299)	-1.686*** (0.289)	-1.704*** (0.280)
Maths Test score	0.0951*** (0.023)	0.0971*** (0.024)	0.0681** (0.021)	0.0646** (0.021)	0.0602** (0.021)	0.0611** (0.021)
Word Reading Test score	0.0489*** (0.005)	0.0502*** (0.005)	0.0484*** (0.004)	0.0456*** (0.004)	0.0437*** (0.004)	0.0440*** (0.004)
Pattern Construction Test score	0.0313*** (0.007)	0.0258*** (0.007)	0.0168* (0.007)	0.0166* (0.007)	0.0172* (0.007)	0.0159* (0.007)
Constant	6.932 (5.809)	34.41*** (7.845)	36.48*** (7.509)	36.02*** (7.417)	35.91*** (7.317)	35.84*** (7.194)
N	829	829	823	823	823	823
R²	0.703	0.737	0.769	0.773	0.775	0.776

Standard errors in parentheses. Reference category in brackets. Coefficients from linear regression model. Ns are unweighted; coefficients are weighted for initial survey design and for attrition to the level of the main wave-four survey.

^a $p < 0.10$, ^{*} $p < 0.05$, ^{**} $p < 0.01$, ^{***} $p < 0.001$

(a) Outcome is summed survey-reported teacher judgement; range: 7–35.

(b) See Table 4 for details of covariates at each specification. Full table of coefficients available from author on request.

Table 6 shows that results hold when teacher judgement of reading ability is considered in isolation (conditional upon children's reading ability test score, maths and PCT test scores, and all non-cognitive test covariates), as well as when maths ability is considered alone. Again, judgements of both reading and maths ability, like summed overall teacher judgements, are related to the stream in which a pupil is situated.

Table 6: Differences in survey-reported teacher judgements of level of reading/maths 'ability and attainment' according to pupils' stream placement (specification six)^{(a)(b)}

	Reading judgement	Maths judgement
Top stream	1.193*** (0.158)	1.143*** (0.158)
(Middle stream)	0 (.)	0 (.)
Bottom stream	-0.837*** (0.170)	-1.087*** (0.182)
Maths Test score	0.00523 (0.011)	0.0499*** (0.012)
Word Reading Test score	0.0338*** (0.002)	0.0102*** (0.002)
Pattern Construction Test score	0.00426 (0.003)	0.0111*** (0.003)
Cut 1: Constant	-10.09** (3.022)	-10.67** (3.485)
Cut 2: Constant	-7.912** (3.015)	-8.587* (3.471)
Cut 3: Constant	-5.563+ (3.015)	-6.198+ (3.507)
Cut 4: Constant	-3.465 (3.027)	-4.219 (3.515)
N	843	839

Standard errors in parentheses. Reference category in brackets. Coefficients from ordered probit models. Ns are unweighted; coefficients are weighted for initial survey design and for attrition to the level of the main wave four survey.

^a $p < 0.10$, ^{*} $p < 0.05$, ^{**} $p < 0.01$, ^{***} $p < 0.001$

(a) Outcomes are survey-reported teacher judgements of reading/maths ability; range: 1–5.

(b) See Table 4 for details of covariates at each specification. Full table of coefficients available from author on request.

Sensitivity checks, alternative specifications and extensions of analyses

As noted in the introduction, previous research suggests that the influence of streaming may manifest itself not just through its effects on teacher perceptions, but through, for example, its impact on pupils' self-esteem, or on educational opportunities. Given that the data used in this study is observational, it is possible that, despite numerous controls, the correlations indicated may to some extent reflect a difference in actual pupil performance, accurately reported by teachers, rather than a bias in teacher perceptions engendered by stream placement. Despite consistent findings across domains of survey-reported teacher judgement, it remains possible that alternative non-observed factors might intervene. So far as possible, within the data available, a number of sensitivity checks therefore continue to interrogate the reliability of results.

The first examines whether removing teachers' reports regarding less 'academic' subjects from the overall survey-reported summed judgement of 'ability and attainment' affects findings. Results are entirely consistent using this alternative outcome. The second replicates analyses without MCS survey weights but with clustering of standard errors at the school level. Again,

findings are consistent and remain significant at the 5 per cent level. Third, linear versions of models where the five levels of response regarding reading and maths ability/assessment are treated as continuous variables yield equivalent results.

As described, MCS survey-reported judgements are the preferred measure for investigation of teacher perceptions, due to their independence from the biasing influences of the school system, and their arguably more direct proxy of a teacher's internal representation of a child. However, a fourth sensitivity check uses Key Stage 1 (KS1) scores (awarded by teachers at the end of the year in which MCS surveying took place) as an alternative depiction of teacher judgement, and finds patterns that are consistently significant at the 5 per cent level and that are in the same direction as results from analyses of the survey measures. Otherwise equivalent pupils who are in the top stream score higher at KS1, and those in the bottom stream score lower. This holds across the average point score, and for reading and maths levels, respectively. See Campbell (2014) for further detail on these analyses and others reported above (though regressions have been modified since this version of analyses, relationships between covariates and outcomes remain substantively unaltered); findings for all alternative specifications and for covariates from all these models are available from the author on request.

A final interrogation seeks further to disentangle relationships with perceptions from alternative routes from stream placement to teacher judgements. These sub-analyses delve further into modelling using the teacher survey responses, dividing the sample into thirds according to the length of time elapsed between cognitive testing and teacher survey.

Behind this strategy is an assumption that individual teacher-level judgement bias related to stream placement will remain approximately stable over time. If there is a greater magnitude of relationship between stream placement and teacher ratings for children with a longer lapse to judgement from the cognitive test baseline, this will suggest that the average estimates found so far for the whole sample may be driven by explanatory factors other than teachers' perceptions: either by alternative routes through which streaming itself may influence progress or by unobserved confounders.

As an example, Table 7 shows coefficients at specification six for pupils with a zero-to-two month lapse between cognitive tests and survey, a three-to-four month lapse, and a lapse of more than five months, respectively. It indicates that the magnitude of the coefficient for the relationship between bottom stream placement and overall teacher judgement appears in fact to lessen when there is a longer lapse, while the coefficient for ratings of children in the top stream peaks in the three-to-four month group, then lessens in cases with a lapse of over five months. Based on the supposition that magnitude of perceptual bias should remain stable over the school year, this suggests that the apparent association between stream placement and teacher ratings is more predominantly due to this bias than to actual changes in pupil performance – because these, in combination with perceptions, would lead to a relationship whose magnitude grows over time.

Analyses by lapse subgroup for teacher judgements of reading and maths proficiency, respectively, also supports this conclusion: here too, there is no overall increase with time. Although sample sizes for all these analyses are relatively small (there is some missing data on teacher survey date), and results should therefore be treated with caution, they provide some further support to the hypothesis that stream placement may have an impact directly upon teacher perceptions. (Again, all additional analyses are available from the author on request.)

Table 7: Difference in summed survey-reported teacher judgement of ‘ability and attainment’ according to pupils’ stream placement: subgroups for cognitive test – teacher survey lag^{(a)(b)}

	Full sample	0–2 month lag	3–4 month lag	5+ month lag
Top stream	2.569*** (0.258)	2.071 *** (0.372)	3.326*** (0.482)	2.408*** (0.549)
(Middle stream)	0 (.)	0 (.)	0 (.)	0 (.)
Bottom stream	-1.704*** (0.280)	-2.039*** (0.525)	-1.521 ** (0.521)	-1.233** (0.395)
Maths Test score	0.0611 ** (0.021)	0.0568 (0.037)	0.0895* (0.037)	0.0995** (0.036)
Word Reading Test score	0.0440 *** (0.004)	0.0566 *** (0.008)	0.0308 *** (0.007)	0.0404 *** (0.008)
Pattern Construction Test score	0.0159* (0.007)	0.0467 *** (0.008)	0.00859 (0.011)	0.00855 (0.013)
Constant	35.84*** (7.194)	56.38** (18.098)	-6.170 (20.725)	43.26 *** (9.883)
N	823	229	291	249
R²	0.776	0.861	0.830	0.799

Standard errors in parentheses. Reference category in brackets. Coefficients from linear regression model. Ns are unweighted; coefficients are weighted for initial survey design and for attrition to the level of the main wave-four survey.

* $p < 0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

(a) Outcome is summed survey-reported teacher judgement; range: 7–35.

(b) See Table 4 for details of covariates at each specification. Full table of coefficients available from author on request.

Discussion

This research set out to explore whether teacher perceptions of pupils are related to the stream to which a child is allocated, and whether this mechanism may therefore provide some explanation for previously evidenced associations between stream placement and children’s progress. Having controlled for recent performance on relevant cognitive tests, as well as a wide range of potential confounding factors, it finds consistent relationships between assigned stream level and subsequent teacher judgements of pupils’ academic ability and attainment. This holds across a variety of specifications and alternative analyses.

The hypothesis that teacher judgements of pupils are influenced by the stream to which pupils are allocated is therefore supported. As there is also evidence that certain groups of pupils (boys, low-income pupils, pupils whose parents have fewer qualifications, summer-born children) are over-represented in lower streams, and under-represented in the highest groupings, streaming is implicated as potentially instrumental in disparities in attainment trajectories. Results here therefore belie the notion that streaming can objectively be implemented as an accurate replication of ‘natural’, permanent, stable, underlying differences between pupils.

Limitations and further research

As noted, because the data used in this paper are observational, it is feasible that alternative factors could explain the patterns described. So far as is possible using the information available, this is addressed. Findings are congruent with previous studies, so the explanation favoured here, of a direct influence of streaming upon teacher perceptions, seems arguably coherent.

However, survey sample data are somewhat limited in their capacity to unpick the processes and complexities of streaming. It is not tenable, for example, to explore completely within the MCS differences in relationships according to teacher characteristics, or other school-level factors. In order to do this, comprehensive, whole-school samples are necessary – and in order for these to be nationally meaningful, they should include as many institutions as possible. Pending data linkage of the MCS to each cohort member's whole-school peer group in the National Pupil Database (NPD) will go some way towards addressing this need.

Collecting information on whether streaming takes place and on the stream placement of each individual pupil, and making this information available for longitudinal analysis through the entire NPD, would also, more fully, allow proper, transparent scrutiny of the impacts of the practice. In addition, current government-funded evaluations of grouping in secondary schools (www.ucl.ac.uk/ioe/departments-centres/centres/groupingstudents/) should be built upon to investigate streaming in early primary school.

Of course, in order definitively and causally to isolate the direct effect of stream placement on teachers' judgements, random allocation to streaming and/or to placement would be necessary. Ostensibly, this appears ethically questionable, given the body of observational research that indicates detrimental effects for those in the lower streams. However, as, to date, the roll-out of streaming has crept unchecked (Hallam and Parsons, 2013), a kind of 'reverse randomized control trial' may be a reasonable and workable possibility. A systematic trial where a sample of primary schools already implementing streaming are identified, then randomly allocated to cessation of the practice for a given period, would create a valid counterfactual comparison group for better causal inference.

Conclusion

This paper has suggested that stream placement can have a relationship with teacher judgements of primary school children's 'ability and attainment'. Given the recent slide back towards early streaming, and the evidence on the importance of teacher judgements, this indication of a route through which streaming may influence progress is immediately applicable to current policy, practice and discussion. Analyses here add to an accumulation of research on the potential effects of streaming, and on plausible mechanisms for these. Use of the practice – especially among very young pupils – should no longer be ignored, assumed to be straightforwardly reasonable, uncomplicatedly rational or reflective of a 'natural' order. Streaming should explicitly be acknowledged by policymakers, and the implications of its use considered by practitioners.

Acknowledgements

This paper presents analysis forming part of my PhD, which I am very grateful to the Economic and Social Research Council for funding. It uses data from the Millennium Cohort Study. I am grateful to the Centre for Longitudinal Studies, UCL Institute of Education for the use of these data, and to the UK Data Archive and Economic and Social Data Service for making them available. However, these organizations bear no responsibility for analysis or interpretation.

Many thanks to Lorraine Dearden, Anna Vignoles and Alissa Goodman for useful comments and suggestions.

Notes on the contributor

Tammy Campbell is a researcher in the Centre for Analysis of Social Exclusion, London School of Economics. Her PhD was awarded by the Department of Quantitative Social Science, UCL Institute of Education, and she worked previously as a government social researcher. Before that, Tammy held various practitioner roles with children and young people.

References

- Alvidrez, J. and Weinstein, R.S. (1999) 'Early teacher perceptions and later student academic achievement'. *Journal of Educational Psychology*, 91 (4), 731–46.
- Ansalone, G. (2003) 'Poverty, tracking, and the social construction of failure: International perspectives on tracking'. *Journal of Children and Poverty*, 9 (1), 3–20.
- Blatchford, P., Hallam, S., Ireson, J., Kutnick, P. and Creech, A. (2010) 'Classes, groups and transitions: Structures for teaching and learning'. In Alexander, R. (ed.) *The Cambridge Primary Review Research Surveys*. London: Routledge, 548–88.
- Boaler, J. (1997) 'Setting, social class and survival of the quickest'. *British Educational Research Journal*, 23 (5), 575–95.
- Boaler, J., Wiliam, D. and Brown, M. (2000) 'Students' experience of ability grouping: Disaffection, polarisation and the construction of failure'. *British Educational Research Journal*, 26 (5), 631–48.
- Bradbury, A. (2011) 'Equity, ethnicity and the hidden dangers of "contextual" measures of school performance'. *Race Ethnicity and Education*, 14 (3), 277–91.
- Brophy, J.E. and Good, T.L. (1970) 'Teachers' communication of differential expectations for children's classroom performance: Some behavioral data'. *Journal of Educational Psychology*, 61 (5), 365–74.
- Burgess, S. and Greaves, E. (2009) *Test Scores, Subjective Assessment and Stereotyping of Ethnic Minorities* (Working Paper 09/221). Bristol: Centre for Market and Public Organisation. Online. www.bris.ac.uk/cmpo/publications/papers/2009/wp221.pdf (accessed 20 May 2016).
- Campbell, T. (2013) *In-School Ability Grouping and the Month of Birth Effect: Preliminary evidence from the Millennium Cohort Study* (CLS Cohort Studies Working Paper 2013/1). London: Centre for Longitudinal Studies. Online. www.cls.ioe.ac.uk/shared/get-file.ashx?itemtype=document&id=1618 (accessed 20 May 2016).
- Campbell, T. (2014) 'Stratified at seven: In-class ability grouping and the relative age effect'. *British Educational Research Journal*, 40 (5), 749–71.
- Campbell, T. (2015) 'Stereotyped at seven? Biases in teacher judgement of pupils' ability and attainment'. *Journal of Social Policy*, 44 (3), 517–47.
- Conservative Party (2007) *Raising the Bar, Closing the Gap: An action plan for schools to raise standards, create more good school places and make opportunity more equal*. London: Conservative Party. Online. <http://image.guardian.co.uk/sys-files/Education/documents/2007/11/20/newopps.pdf> (accessed 20 May 2016).
- Croizet, J.-C. and Claire, T. (1998) 'Extending the concept of stereotype threat to social class: The intellectual underperformance of students from low socioeconomic backgrounds'. *Personality and Social Psychology Bulletin*, 24 (6), 588–94.
- DfES (Department for Children, Schools and Families) (2008) *21st Century Schools: A world-class education for every child*. Nottingham: Department for Children, Schools and Families. Online. www.education.gov.uk/consultations/downloadableDocs/21st%20Century%20Schools%20consultation%20document.doc (accessed 20 May 2016).
- DfE (Department for Education) (1992) *White Paper: Choice and diversity: A new framework for schools*. London: HMSO. Online. www.educationengland.org.uk/documents/wp1992/choice-and-diversity.html (accessed 20 May 2016).

- DfE (Department for Education) (2010) *The Importance of Teaching: The Schools White Paper 2010*. Norwich: TSO. Online. www.gov.uk/government/publications/the-importance-of-teaching-the-schools-white-paper-2010 (accessed 20 May 2016).
- DfE (Department for Education) (2016) *Educational Excellence Everywhere*. London: HMSO. Online. www.gov.uk/government/uploads/system/uploads/attachment_data/file/508447/Educational_Excellence_Everywhere.pdf (accessed 20 May 2016).
- DfES (Department for Education and Skills) (2005) *Higher Standards, Better Schools for All: More choice for parents and pupils*. Norwich: TSO. Online. www.educationengland.org.uk/documents/pdfs/2005-white-paper-higher-standards.pdf (accessed 20 May 2016).
- Dunne, M., Humphreys, S., Sebba, J., Dyson, A., Gallannaugh, F. and Muijs, D. (2007) *Effective Teaching and Learning for Pupils in Low Attaining Groups* (Research Report DCSF-RR011). Nottingham: Department for Children, Schools and Families. Online. <http://dera.ioe.ac.uk/6622/1/DCSF-RR011.pdf> (accessed 20 May 2016).
- Education Endowment Foundation (2017) Setting or Streaming: Negative impact for very low or no cost, based on moderate evidence. Online. [https://educationendowmentfoundation.org.uk/pdf/toolkit/?id=127&t=Teaching%20and%20Learning%20Toolkit&e=127&s=](https://educationendowmentfoundation.org.uk/pdf/generate/?u=https://educationendowmentfoundation.org.uk/pdf/toolkit/?id=127&t=Teaching%20and%20Learning%20Toolkit&e=127&s=) (accessed 30 August 2017).
- Francis, B., Archer, L., Hodgen, J., Pepper, D., Taylor, B. and Travers, M.-C. (2017) 'Exploring the relative lack of impact of research on "ability grouping" in England: A discourse analytic account'. *Cambridge Journal of Education*, 47 (1), 1–17.
- Good, T.L. (1987) 'Two decades of research on teacher expectations: Findings and future directions'. *Journal of Teacher Education*, 38 (4), 32–47.
- Hallam, S., Ireson, J., Lister, V., Andon Chaudhury, I. and Davies, J. (2003) 'Ability grouping practices in the primary school: A survey'. *Educational Studies*, 29 (1), 69–83.
- Hallam, S. and Parsons, S. (2013) 'Prevalence of streaming in UK primary schools: Evidence from the Millennium Cohort Study'. *British Educational Research Journal*, 39 (3), 514–44.
- Hansen, K. (ed.) (2012) *Millennium Cohort Study First, Second, Third and Fourth Surveys: A guide to the datasets*. 7th ed. London: Centre for Longitudinal Studies. Online. www.cls.ioe.ac.uk/shared/get-file.ashx?id=598&itemtype=document (accessed 20 May 2016).
- Hansen, K. (2016) 'The relationship between teacher perceptions of pupil attractiveness and academic ability'. *British Educational Research Journal*, 42 (3), 376–98.
- Hansen, K. and Jones, E.M. (2011) 'Ethnicity and gender gaps in early childhood'. *British Educational Research Journal*, 37 (6), 973–91.
- Harlen, W. (2007) *The Quality of Learning: Assessment alternatives for primary education* (Primary Review Research Survey 3/4). Cambridge: Primary Review.
- Higgins, S., Katsipataki, M., Kokotsaki, D., Coleman, R., Major, L.E. and Coe, R. (2014) *The Sutton Trust-Education Endowment Foundation Teaching and Learning Toolkit*. London: Education Endowment Foundation.
- Huang, Y. and Gatenby, R. (2010) *Millennium Cohort Study Sweep 4 Teacher Survey: Technical report*. London: Centre for Longitudinal Studies. Online. www.cls.ioe.ac.uk/shared/get-file.ashx?id=489&itemtype=document (accessed 20 May 2016).
- Ireson, J. and Hallam, S. (1999) 'Raising standards: Is ability grouping the answer?'. *Oxford Review of Education*, 25 (3), 343–58.
- Johnson, J., Rosenberg, R., Platt, L. and Parsons, S. (2011) *Millennium Cohort Study Fourth Survey: A guide to the teacher survey dataset*. London: Centre for Longitudinal Studies. Online. www.cls.ioe.ac.uk/shared/get-file.ashx?id=1341&itemtype=document (accessed 20 May 2016).
- Kutnick, P., Hodgkinson, S., Sebba, J., Humphreys, S., Galton, M., Steward, S., Blatchford, P. and Baines, E. (2006) *Pupil Grouping Strategies and Practices at Key Stage 2 and 3: Case studies of 24 schools in England*. London: Department for Education and Skills. Online. www.academia.edu/11198384/Pupil_Grouping_Strategies_and_Practices_at_Key.Stage_2_and_3_Case_Studies_of_24_Schools_in_England (accessed 20 May 2016).
- Kutnick, P., Sebba, J., Blatchford, P., Galton, M. and Thorp, J. (2005) *The Effects of Pupil Grouping: Literature review* (Research Report RR688). Nottingham: Department for Education and Skills. Online. <https://core.ac.uk/download/files/161/15171799.pdf> (accessed 20 May 2016).

- Miller, K. and Satchwell, C. (2006) 'The effect of beliefs about literacy on teacher and student expectations: A further education perspective'. *Journal of Vocational Education and Training*, 58 (2), 135–50.
- Mostafa, T. (2013) *Millennium Cohort Study Technical Report on Response in the Teacher Survey in MCS 4 (Age 7)*. London: Centre for Longitudinal Studies. Online. www.cls.ioe.ac.uk/shared/get-file.ashx?id=1749&itemtype=document (accessed 29 July 2017).
- OECD (Organisation for Economic Co-operation and Development) (2012) *Equity and Quality in Education: Supporting disadvantaged students and schools*. Paris: OECD Publishing. Online. www.keepeek.com/Digital-Asset-Management/oecd/education/equity-and-quality-in-education_9789264130852-en (accessed 20 May 2016).
- Parsons, S. and Hallam, S. (2014) 'The impact of streaming on attainment at age seven: Evidence from the Millennium Cohort Study'. *Oxford Review of Education*, 40 (5), 567–89.
- Reay, D. (2006) 'The zombie stalking English schools: Social class and educational inequality'. *British Journal of Educational Studies*, 54 (3), 288–307.
- Reeves, D.J., Boyle, W.F. and Christie, T. (2001) 'The relationship between teacher assessments and pupil attainments in standard test tasks at Key Stage 2, 1996–98'. *British Educational Research Journal*, 27 (2), 141–60.
- Rosenthal, R. and Jacobson, L. (1968) 'Pygmalion in the classroom'. *Urban Review*, 3 (1), 16–20.
- Rubie-Davies, C.M. (2010) 'Teacher expectations and perceptions of student attributes: Is there a relationship?'. *British Journal of Educational Psychology*, 80 (1), 121–35.
- Rubie-Davies, C.M., Weinstein, R.S., Huang, F.L., Gregory, A., Cowan, P.A. and Cowan, C.P. (2014) 'Successive teacher expectation effects across the early school years'. *Journal of Applied Developmental Psychology*, 35 (3), 181–91.
- Shih, M., Pittinsky, T.L. and Trahan, A. (2005) *Domain Specific Effects of Stereotypes on Performance* (Faculty Research Working Paper RWPO5-026). Cambridge, MA: John F. Kennedy School of Government, Harvard University. Online. www.cs.cmu.edu/~cfrieze/courses/Shih.pdf (accessed 20 May 2016).
- Slavin, R.E. (1990) 'Achievement effects of ability grouping in secondary schools: A best-evidence synthesis'. *Review of Educational Research*, 60 (3), 471–99.
- Steele, C.M. and Aronson, J. (1995) 'Stereotype threat and the intellectual test performance of African Americans'. *Journal of Personality and Social Psychology*, 69 (5), 797–811.
- Thomas, S., Madaus, G.F., Raczek, A.E. and Smees, R. (1998) 'Comparing teacher assessment and standard task results in England: The relationship between pupil characteristics and attainment'. *Assessment in Education: Principles, Policy and Practice*, 5 (2), 213–46.
- University of London. Institute of Education. Centre for Longitudinal Studies, *Millennium Cohort Study: Fourth Survey, Teacher Survey, 2008* [computer file]. Colchester, Essex: UK Data Archive [distributor], August 2011a. SN: 6848, <http://dx.doi.org/10.5255/UKDA-SN-6848-1>.
- University of London. Institute of Education. Centre for Longitudinal Studies, *Millennium Cohort Study, 2001–2008: Linked Education Administrative Dataset, England: Secure Access* [computer file]. Colchester, Essex: UK Data Archive [distributor], November 2011b. SN: 6862, <http://dx.doi.org/10.5255/UKDA-SN-6862-2>.
- University of London. Institute of Education. Centre for Longitudinal Studies, *Millennium Cohort Study: Fourth Survey, 2008* [computer file]. 4th Edition. Colchester, Essex: UK Data Archive [distributor], December 2012a. SN: 6411, <http://dx.doi.org/10.5255/UKDA-SN-6411-3>.
- University of London. Institute of Education. Centre for Longitudinal Studies, *Millennium Cohort Study: Third Survey, 2006* [computer file]. 6th Edition. Colchester, Essex: UK Data Archive [distributor], December 2012b. SN: 5795, <http://dx.doi.org/10.5255/UKDA-SN-5795-3>.
- William, D. and Bartholomew, H. (2004) 'It's not which school but which set you're in that matters: The influence of ability grouping practices on student progress in mathematics'. *British Educational Research Journal*, 30 (2), 279–93.
- Yopyk, D.J.A. and Prentice, D.A. (2005) 'Am I an athlete or a student? Identity salience and stereotype threat in student-athletes'. *Basic and Applied Social Psychology*, 27 (4), 329–36.